学苑・人間文化学科特集 No.773 (10)～(21) (2005・3)

Applying the Rasch Measurement Theory to Linking Graded Test Forms[1]

Kay Irie

Abstract

It has been a difficult task for teachers to compare the abilities of students from different grades or levels using traditional testing theory (TTT) because all students must take a common test that includes items covering the entire range of abilities. However, by using Rasch measurement theory (RMT), it is possible to give graded forms of the test to each group of students by including some overlapping items. This study provides a brief introduction to RMT followed by a demonstration of an application of the theory to measure abilities of students from different grades in junior high schools. The process of equating three different English achievement test forms, one for each grade administered to a total of 900 junior high school students is described. A list of useful resources on RMT is also provided in the appendix.

Introduction

One of the challenges for educators, including language teachers, is to fairly assess students' performance across different years or levels in the same course with consistency. Unless the difficulty of tests for the same course over years or across sections is ensured, students at the same level of performance or achievement may pass or fail, depending on when they took the course or which section they happened to be assigned to. Moreover, in Japan, keeping the difficulty level constant is of great importance in high school and university entrance examinations. Using traditional testing theory (TTT), despite the tireless effort exerted by teachers to maintain a certain level of difficulty across sections or years, students' performance cannot be directly compared based on the raw scores. Equating of two test forms is possible in TTT only if all students take both forms of the test, which is often logistically difficult. The conversion of the raw scores to *z*-scores can be also problematic as it works only on an assumption that the test takers are the same people.

How can we compare the students who have taken different forms of a test or different tests for the same course? In the field of educational testing, a modern test theory (item response theory = IRT) for measuring students' abilities and achievements has been developed to meet this challenge of equating tests. Probability theory allows us to compare students' performances regardless of differences in their abilities and test items used: It is possible to standardize scores of two or more tests as long as they share common items that provide

sufficient linkage between them. In fact, this is the principle used by Educational Testing Service (ETS) to calibrate standardized scores of all test takers across different forms of TOEFL or TOEIC tests. In Japan, it is used in Examination for Japanese University Admission for International Students (EJU), a standardized test to evaluate the Japanese language skills and the basic academic abilities of non-native speakers of Japanese who wish to study at higher education institutions in Japan.

Among the several models of IRT, Rasch measurement theory (RMT, or one-parameter IRT), has been gaining acceptance in educational institutions. In the United States, the University System of Georgia has been using the model to assess students' reading and writing skills across the universities since 2002. The model has also been applied to a district-wide assessment of high school students' performances in content areas in Glendale, Arizona. In Singapore, the National Institute of Education (NIE) has developed a nationwide standardized comprehensive English test known as NIE Computerized English Language Test (NIECELT) using RMT.

Although the changes in educational measurement are taking place on one level, TTT is still dominantly used to equate tests by most institutions and teachers by simply relying on their expertise to maintain the same level of difficulty. One obstacle dissuading testers from trying out this modern test theory is that the theory involves sophisticated mathematical formulas that might be difficult for teachers to grasp. However, with the recent publication of introductory books on RMT (e.g. Bond & Fox, 2001; McNamara, 1996) and the development of computer technology and accessible software, the application of the model has become much easier.

The purpose of this paper is simply to demonstrate how RMT was applied to equate different forms of a multiple-choice test by following the steps taken to develop an English achievement test for junior high school students. After a brief introduction to RMT principles and the background of the test, the process involved in equating test forms of differing difficulty levels will be explained.

RMT and Equating

RMT is a mathematical framework in which raw data can be transformed into abstract, equal-interval scales (Bond & Fox, 2001). Equal intervals are formed through log transformation of raw data odds, that is, the test-takers' probability of getting correct answers. A series of probabilistic equations enables the abstraction of the scale. Moreover, in RMT, an objective scale can be constructed without being influenced by the ability of the persons it measures, unlike in classical testing theory (Bond & Fox, 2001, p. 7; for the formulas, see Wright & Masters, 1982; Wright & Stone, 1979).

Equating is a process by which measures from various test forms, all measuring the same construct are located on the same linear scale. Although equating is possible in classical testing

theory, using the equipercentile equating approach, all the tests must be taken by the same group of people (Styles & Andrich, 1993; Wolfe, 2000). On the other hand, in RMT, the zero point can be set anywhere on the scale and the scales are independent of test-takers' abilities. Taking advantages of such attributes of RMT, common items can link different test forms and result in measures that are placed on a single scale.

Achievement Test

The test was designed and developed as part of my dissertation research on language learning motivation (L2 motivation) of Japanese junior high school students (Irie, 2005). The study examined changes in their motivation through a large scale cross-sectional study (N = 979 drawn from Schools H, K, and N) and a three-year longitudinal study following a cohort of students (N = 76) at school H. In order to compare the mastery levels of all sample groups in the cross-sectional study and the level of the cohort group at different points of time in the longitudinal study, a unified achievement measure was required. Although RMT was used to assess the progress (changes) in the mastery level of the longitudinal cohort and the scores from both the cross-sectional and longitudinal studies were consequently equated on the same scale, the focus of the present study is limited on the process of equating test forms of different test levels (graded test forms) for the cross-sectional study.

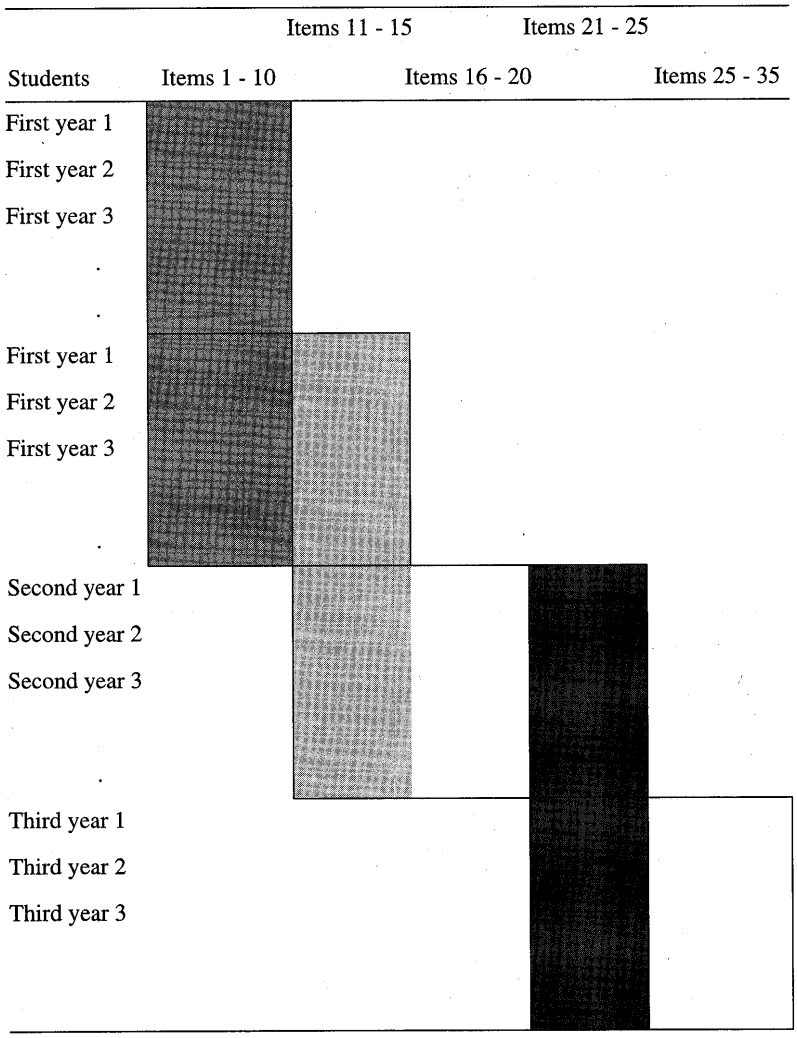
In the pilot studies, a three-paragraph cloze test to assess the mastery level of junior high school English was developed, but it turned out to be problematic for the main studies for two reasons. One was that it would be too difficult for new first year students. The baseline achievement measure in the longitudinal study had to be taken before the participants at School H, where the longitudinal study was conducted, actually began learning English. For these students, a paragraph of several sentences could be overwhelming. Their English teacher was concerned about the possible negative impact of an unnecessarily difficult and long test for 12-year-old beginning learners. Thus, a gap-filling multiple-choice format was developed for the main studies (see Appendix A). Another problem was that the cloze test was too long. The time that could be allocated for the achievement test was 15 minutes maximum. At the same time, the test needed to have a sufficient number of items. To spread out the participants enough to ensure a normal distribution, the grammatical points had to be selected carefully from the first, second, and third year materials. To solve the dilemma between time and content constraints, I decided to create four short forms of the test and link them by common item equating based on RMT (Bond & Fox, 2001; McNamara, 1996; Wolfe, 2000; Wright & Stone, 1979). In other words, a multiple-choice gap-filling test (see Appendix A) to measure achievement over the three years was divided into four overlapping segments.

This use of RMT, called *vertical equating*, then makes it possible to place the students from different grades onto the same scale by giving difficult items to the third year students, middle-

range items to the second year students, and easy items to the first year students. This approach was suitable for the study, because it saved the first year students from facing questions targeted to measure the ability of third year students, and each test form was short enough to be completed in 15 minutes.

Design of the Achievement Test

Four multiple-choice gap-filling test forms with five to ten overlapping items were constructed in order to measure the participants' mastery of the materials covered in three years: Form A for the newly entered first year students in April, Form B1 for those at the end of the first year, Form B2 for those at the end of second year, and Form B3 for those at the end of the third year (see Figure 1). Items 1 to 35 should gradually increase in difficulty, from the left to the right on Figure 1.



Note. The size of boxes does not correspond with the number of items.

Figure 1. Outline of the four achievement test forms.

Table 1 shows the actual linking of the four forms: Form A was used with the first year students at School H in April, 2000 to obtain the baseline data in the longitudinal study; Forms B1, B2, and B3 were used with the first, second, and third year students in both the cross-sectional and longitudinal studies.. The items in Form A were drawn from phrases typically introduced aurally/orally in children's English textbooks and lessons. Easy-to-read words and obviously wrong distracters were chosen so that even if the students had not formally learned to read, they would be more likely to find the correct answers as long as they were familiar with the expressions. To measure the progress of the first year students, I confirmed that these items were introduced in both textbooks, *New Crown* and *One World*, during the first year in junior high school. The additional five items in Form B1 and the items included in Forms B2 and B3 were not directly taken from the textbooks, but they were within the range of grammar, vocabulary, and conversational expressions covered by them. In RMT, a person's ability cannot be estimated if they obtain a perfect or a zero score. Therefore, in order to decrease the possibility of perfect scores, a few items from the senior high school level were also included in Forms B2 and B3.

Table 1 The Gap-filling Test Forms

	A	B1	B2	B3	B1'	B2'	B3'
Items							
1	1	1	*	*	1	*	*
2	2	2	*	*	*	*	*
3	3	3	*	*	2	*	*
4	4	4	*	*	3	*	*
5	5	5	*	*	*	*	*
6	6	6	*	*	*	*	*
7	7	7	*	*	*	*	*
8	8	8	*	*	4	1	*
9	9	9	*	*	5	2	*
10	10	10	*	*	6	3	*
11	*	11	1	*	7	4	*
12	*	12	2	*	8	5	*
13	*	13	3	*	9	6	*
14	*	14	4	*	10	7	*
15	*	15	5	*	11	8	*
16	*	*	6	*	*	9	1
17	*	*	7	*	*	10	2
18	*	*	8	*	*	11	3
19	*	*	9	*	*	12	4
20	*	*	10	*	*	13	5
21	*	*	11	1	*	14	6

	A	B1	B2	B3	B1'	B2'	B3'
Items							
22	*	*	12	2	*	15	7
23	*	*	13	3	*	16	8
24	*	*	14	4	*	*	*
25	*	*	15	5	*	17	9
26	*	*	*	6	*	18	10
27	*	*	*	7	*	*	11
28	*	*	*	8	*	*	12
29	*	*	*	9	*	*	13
30	*	*	*	10	*	*	14
31	*	*	*	11	*	*	*
32	*	*	*	12	*	*	15
33	*	*	*	13	*	*	16
34	*	*	*	14	*	*	*
35	*	*	*	15	*	*	17
36 New					*	*	18
37 New					14	19	*
38 New					15	20	*
39 New					*	*	19
40 New					*	*	20
41 New					12	*	*
42 New					13	*	*

Revising the Test

In April 2000, Form A was administered to the first year students at School H for both the cross-sectional and longitudinal studies. In March 2001, a year later, Form B1 was administered to the same group of students, Form B2 to the second year students, and Form B3 to the third year students, all at School H. As there was a year between the administration at School H and the next administration at School K and N, I took advantage of the interval and examined the reliability and fit statistics of persons and items of the School H data. My intention was to confirm that the 35 items were functioning properly before collecting more data at Schools K and N.

In RMT, the reliability of the estimation of item difficulty and person ability are expressed in terms of an *item reliability index* and *person reliability index*. Both indicate whether there is enough spread of item difficulty and person ability (Bond & Fox, 2001). The item reliability index shows the replicability of item placement along the continuum if these same items were taken by a different group of people. Likewise, the person reliability index shows how confident we can be about the person ordering, that is, how much we can expect the same ordering of persons, if we administer a different set of items measuring the same construct. The reliability indices are on a scale of 0 to 1, with values close to 1 suggesting high reliability. Rasch analysis also provides fit statistics, which are useful in evaluating the degree to which the data fit the Rasch model. This was also examined in order to determine whether each item fits construct (item fit), and whether each person deviated only within an acceptable range of the expected ability pattern (person fit). There are two groups of fit statistics, *mean square* values and *t*-distribution, both of which are presented as *outfit* and *infit* statistics.

According to Bond and Fox (2001), outfit is the sum of squared standardized residuals. Infit is an information-weighted sum which weights more on the well-targeted observations than extreme observations. Thus, the infit statistics are more sensitive to irregular patterns in the response string than the outfit statistics, which reflect the responses of persons whose estimated ability is well above or below the item's difficulty. Mean square values, the unstandardized fit statistics, are the average of the squared residuals for each item. The standardized fit statistics, the mean square values are expressed as *t*-values in the *t*-distribution. The residual values indicate the discrepancy between the ideal performance by the Rasch model and the observed performance of the item. To evaluate the fit statistics, I followed the rule of thumb suggested by McNamara (1996, p. 173): mean square values should range between .75 and 1.3, and *t*-values should range between −2 and +2. When infit and outfit values are above that range, the items or persons are typically viewed as underfitting the Rasch model. When the fit values are below the range, the item or the person is considered to overfit the model.

To examine whether the four forms were sufficiently linked, the four sets of data, A1 and B1 for the School H first year students in April and March respectively, B2 for the School H

second year students, and B3 for School H third year students, were transformed to Rasch logits with item difficulties of the common items anchored. Anchoring means holding the item calibrations for the overlapping common items (the shaded areas in Figure 1) constant across the test forms. The overall fit statistics were within the acceptable range and the item reliability index ranged around .90. However, person reliability ranged between .26 and .36. It increased to around .60 for each test form when the scores were transformed independently of each other. I interpreted this change as an indication of some items that were not working as reliable linking items. To improve the reliability, I increased the number of items in Forms B2 and B3 from 15 to 20 and the number of linking items from five to 10 among the B-form variations.

As the number of items was limited, it was desired that items 1 to 35 gradually increase in difficulty in order to separate the abilities of the participants and achieve high reliability in person ordering. Although the overall fit statistics were all within the acceptable range, the following items were rewritten to improve the overall quality of the instrument. In Forms A and B1, Items 5 and 6 were misfitting with Infit means square values of 1.47 and 1.53 respectively. Items 2 and 7 were acceptable, when only the raw data of Form A was transformed. However, when the items were taken again a year later by the same first year students, all the students got these items correct. As stated above, in RMT, item difficulty and person ability cannot be estimated from perfect scores. Thus, Items 2 and 7 were considered to be too easy to measure the abilities of the first year students in March. Furthermore, item 24 on Form B2, and Items 31 and 34 on Form B3 were misfitting with Infit mean square values of 1.58, 1.45, and 1.67, respectively. Reruns of the analysis without these items affected neither the overall fit nor the item or person reliabilities. I speculated that this was partly due to the smaller number of items canceling out the positive effect of eliminating the problematic items.

To rewrite the above items, I consulted with an experienced English teacher at School H, who collaborated with me in creating the first set of 35 items. She helped me select new items that would fit and discriminate the students' abilities more effectively. Forms B1, B2, and B3 were revised by replacing these problematic items. Table 1 shows the revised forms (B1', B2', and B3'). In short, the total number of items increased from 35 to 42, and the number of forms from four to seven, all of which were linked directly or indirectly through the overlapping items. The measurements are on the same logit scale of item difficulty and person ability as set by the difficulties of the first 10 items in Form A.

Calibrating ability estimates and item difficulty

In order to submit the gap-filling test responses to Rasch analysis using QUEST Version 1 (Adams & Khoo, 1994), first, the participants' answers (1, 2, 3, or 4) were entered in order of the item serial numbers from 1 to 42, and asterisks were used to fill the items that were not included in each form. Unanswered items were interpreted as unattempted missing data and coded. Using items with difficulty estimates that had been established in the longitudinal study

as anchoring items, the ability of the participants in the cross-sectional study was calibrated using QUEST. The item difficulty estimates of Items 1 to 15, which were calibrated from the responses of the longitudinal study participants when they were completing their first year, were used to estimate the ability of all the first year participants in the cross-sectional study. The ability estimates of the second and third year students were obtained through the same process using the item difficulties from the end of the second and third year data in the longitudinal study[2].

The descriptive statistics for achievement scores indicated that the distributions were somewhat peaked in the first year students' data and in the whole sample (Table 2). However, the *z*-scores of these two groups were still within the acceptable range, and a visual inspection of the histograms against the normal curve revealed that they were not adversely leptokurtic. The mean achievement logit score was .78 for the first year students, 1.29 for the second year students, and 1.81 for the third year students. As these logit scores are on a single scale, the items in the three achievement test forms were appropriately graded in terms of difficulty, and the level of achievement rose according to the school year. Table 3 shows that the internal consistency of the achievement test scores (item separation) was .98, indicating that we can be confident about the difficulty order of the items. The person separation index, which is often considered analogous to Cronbach's alpha in classical testing theory was .67. The reason for somewhat low reliability was likely to be due to the large sample size, which resulted in many students being indistinguishable on the logit scale. Both infit/outfit mean squares and infit/outfit *t*-distribution were satisfactory, indicating that the deviation of the observed responses from the

Table 2 Cross-Sectional Study: Achievement Test Descriptive Statistics

Participants	Range	Min.	Max.	*M*	*SD*	Skew.	*SES*	Kurt.	*SEK*
First year[a]	9.67	−5.76	3.91	.78	1.44	−1.13	.14	1.91	.28
Second year[b]	8.58	−3.12	5.46	1.29	1.14	−.09	.13	.67	.27
Third year[c]	6.05	−.59	5.46	1.81	.89	.41	.14	.86	.28
Whole sample[d]	11.22	−5.76	5.46	1.29	1.24	−.86	.08	2.63	.16

Note. [a]$n=30$. [b]$n=334$. [c]$n=303$. [d]$N=937$.

Table 3 Cross-Sectional Study: Achievement Test Reliability Indices

				Mean fit statistics			
				Infit	Outfit		
	M	*SD*	Reliability	*MNSQ*	*MNSQ*	Infit *t*	Outfit *t*
Item[a]	.18	2.08	.98	1.13	1.30	.69	.70
Person[b]	.40	1.09	.67	1.07	1.19	.23	.37

Note. [a]$k=42$. [b]$n=940$.

expected model were acceptable.

Conclusion

In this introduction to RMT, I have outlined the process in which graded test forms are linked and scores are equated on a single scale. The application of RMT equating enabled the comparison of first year students' and third year students' performances without forcing the first year students to struggle with items that were targeted at the third year students. By using this approach, it is possible to compare students' performances and draw consistent cut-off lines for grading across sections or years.

The application of RMT requires the use of computer software. QUEST (Adams & Khoo, 1994), the software used for the present study is suitable for use in a DOS or Mac system. There is a variety of software available for applying RMT to measurements not only in education but the human sciences at large. For Windows users, WINSTEPS (Linacre, 2003) is available. The list of software is provided in Appendix B.

The concept of estimating ability and difficulty may seem strange to those who are used to assessing students' achievement and ability by tallying the number of correct items, which has been the accepted approach in education. However, it is not difficult to see that this approach does not assure equal intervals between 0 and 1, and 99 and 100. The score of 0 does not mean that the student has no ability or achievement whatsoever. Despite the fact that RMT is still in the process of development and refinement, educators should be aware of it as an alternative approach to educational measurements which helps us to overcome such fundamental problems inherent in the current measurements based on TTT.

Notes

[1] Some parts of this paper were written for and included in Irie (2005).

[2] The process of estimating item difficulty and person ability in the longitudinal study is different from equating different tests as it requires a complex process of assessing stability of linking items that is beyond the scope of this paper. A detailed description is provided in Irie (2005).

References

Adams, R., J., & Khoo, S. T. (1994). QUEST (Version Standard) [interactive test analysis system]: A.C.E.R.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum associates, Inc.

Irie, K. (2005). Stability and flexibility of language learning motivation: a multimethod study of Japanese junior high school students. Unpublished doctoral dissertation, Temple University, Philadelphia/Tokyo.

Linacre, J. M. (2003). WINSTEPS [Computer Program]. Chicago: MESA Press.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Styles, I., & Andrich, D. (1993). Linking the standard and advanced forms of the Raven's progressive matrices in both the pencil-and-paper and computer-adaptive-testing formats. *Educational and Psychological Measurement, 53*(4), 905-925.

Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement, 1*(4), 409-434.

Wright, B. D., & Masters, G. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch Measurement*. Chicago, IL: MESA Press.

APPENDIX A

Achievement Test Items

#	Item	A	B1	B1'	B2	B2'	B3	B3'
1	(1. Lemon 2. The 3. Fish 4. This) is a book.	○	○	○				
2	My name is (1. apple 2. dog 3. Tom 4. sun)	○	○					
3	(1. They 2. What's 3. Box 4. New) your name?	○	○	○				
4	I (1. hat 2. new 3. am 4. banana) happy.	○	○	○				
5	(1. Is 2. A 3. How 4. The) are you?	○	○					
6	(1. You 2. Orange 3. Pencil 4. My) like dogs.	○	○					
7	(1. Do 2. My 3. A 4. This) you play tennis?	○	○					
8	I (1. it 2. hello 3. man 4. can) jump.	○	○	○		○		
9	She is my (1. do 2. look 3. you 4. friend).	○	○	○		○		
10	He's 13 years (1. no 2. that 3. you 4. old).	○	○	○	○	○		
11	John visited his aunt (1. Sunday 2. yesterday 3. voyage 4. city).		○	○	○	○		
12	Did you (1. enjoy 2. saw 3. go 4. listen) the concert?		○	○	○	○		
13	There aren't (1. too 2. a 3. any 4. so) clouds in the sky.		○	○	○	○		
14	What day is it? It's (1. Monday 2. March 1st 3. today 4. fine).		○	○	○	○		
15	(1. We 2. These apples 3. It 4. The cake) taste great.		○	○	○	○		
16	What (1. time is 2. time is it 3. is time 4. time it) now?			○				
17	Are you a Yomiuri Giants fan? No, (1. you aren't. 2. you are 3. I'm not. 4. I am.)			○				
18	As (1. soon 2. different 3. far 4. possible) I know, Yohei can not swim.			○		○		
19	(1. What 2. When 3. How 4. Where) does your mother go to work? By train.			○		○		

#	Item	A	B1	B1'	B2	B2'	B3	B3'
20	What's (1. matter 2. bad 3. sick 4. wrong)? I have a headache.				○	○	○	○
21	Ken will teach (1. me English 2. English me 3. to English me 4. to me English).				○	○	○	○
22	My hair is longer than (1. Yoko 2. her 3. hers 4. mine).				○	○	○	○
23	I love ice cream, (1. especially 2. the best 3. more 4. almost) chocolate.				○	○	○	○
24	Are (1. they 2. there 3. these 4. those) any letters for me?				○	○	○	○
25	The post office is (1. next to 2. there 3. from 4. instead) a museum.				○	○	○	○
26	Are you (1. soon 2. will 3. happened to 4. invited) to the party?				○	○	○	○
27	(1. What 2. Who 3. Why 4. While) did you walk all the way to school?				○	○	○	○
28	How (1. much 2. early 3. late 4. long) have you been here?				○		○	
29	It started raining. We (1. should 2. better 3. can 4. not) have brought umbrellas.				○	○	○	
30	Yuta has (1. to 2. is 3. very 4. been) busy for a month.				○		○	○
31	Linda says (1. which 2. if 3. when 4. that) she likes Japan.				○		○	○
32	Is French too difficult to (1. language 2. heard 3. memory 4. learn)?				○		○	○
33	This isn't the (1. crop 2. atomic 3. temple 4. where) we visited before.				○		○	○
34	My brother and I (1. likes each other. 2. alike. 3. look like each other. 4. are each other.)				○			○
35	(1. Do 2. Will 3. You 4. Be) careful, or you will hurt yourself.						○	
36	I'll tell you a story if you (1. early 2. take 3. last 4. are able) a seat.						○	○
37	It's (1. an 2. no 3. an 4. to) use making an excuse for this.						○	○

#	Item	A	B1	B1'	B2	B2'	B3	B3'
38	(1. What 2. Which 3. Where 4. How) do you like Italian food? I love it!						○	○
39	It's getting cold. Why (1. don't 2. do 3. are 4. should) we go inside?						○	○
40	Akio walked as fast as (1. able 2. can 3. possible 4. him).						○	○
41	Would you like (1. eat something? 2. eating something? 3. to eat? 4. something eating?)							○
42	Do you know why (1. Yumi got angry? 2. did Yumi get angry? 3. Yumi get angry? 4. Yumi being angry?)							○
	Number of items	10	15	15	20	20	20	20

Note. ○ indicates the item is included in the form.

APPENDIX B

References on the Rasch Model

Web Sites on Rasch Measurement Theory

The major Web site for Rasch measurement organizations and practitioners: http://www.rasch.org/

Rasch discussion listserv associated with ACER (The Australian Council for Educational Research): http://www.rasch.org/rmt/index.htm

Software

Winsteps for Windows	http://www.winsteps.com/
RUMM for Windows	http://www.rummlab.com.au/
Quest for Mac/DOS	http://www.assess.com/
XCALIBRE (IRT) for Windows	http://www.assess.com/

（入江　恵　　人間文化学科）